EMPLOYEE ATTRITION PREDICTION USING MACHINE LEARNING ALGORITHMS

Lekan AKINODE & Olufunke BADA
Department of Computer Science, Federal Polytechnic, Ilaro, Ogun state, Nigeria
john.akinode@federalpolyilaro.edu.ng; funke.Bada@federalpolyilaro.edu.ng

Abstract

Employees are the most important asset to every organisation. They play critical role in every organisation as they help to generate innovative solutions for the overall development. Employee attrition is regarded as a well-known issue that requires the administration to make the best choices in order to retain highly competent staff. It has indeed been viewed as a crucial concern for businesses due to the negative impact it has on workplace motivation and productivity. This research work compares supervised machine learning methods including Logistic Regression, Naive Bayes Classifier, Random Forest Classifier, and XGBoost in order to forecast an employee's position within an organization. Our research was evaluated using the IBM analytics unbalanced dataset (IBM ,2021) which includes 35 attributes for 1470 employees. Pre-processing was carried out to remove irrelevant features. We created a balanced version from the original dataset in order to obtain results that were realistic. Models are evaluated using evaluation metrics in order to find the most effective model given the available data. Hyperparameter optimization was applied to further improve the results. Finally, cross-validation was used to properly assess our effort. The software was developed using python and anaconda. After reviewing the data, it was discovered that the XGBoost classifier has a better accuracy that is higher than the other techniques, at 85.5 percent, followed by the Random Forest classifier at 84 percent. Precision is higher for the XGBoost classifier (0.72), then for the Random Forest classifier (0.67). The decision tree classifier came in second with an F1 Score of 0.36, while the XGBoost classifier achieved the best F1 Score of 0.40. The XGBoost classifier showed a greater performance in area under the curve (AUC) with 0.50 in the ROC curve. The prediction analysis using the original data set for XGBoost was 85.5%, however it was about 84.6% using a synthetic dataset.

Keywords: Attrition, Employee, Machine Learning, Algorithms

Introduction

The most valuable resources of any company are its employees (Katrina, 2020). They are regarded as the backbone of any organisation. Organizations depend solely on them because they have great influence over the customer relationships and sales. Employees play critical role in organizations, they contribute in generating innovative solutions and thereby enhancing the company's growth and value (Suvoj& Lydia, 2022). Lack of employee recognition can have adverse effect on their motivation; it can lead to loss of interest which makes them feel undervalued (Suvoj& Lydia, 2022). Furthermore, low –self-esteem will affect productivity which in turn affects the organization's overall development (Katrina, 2020). This results to employee attrition, where employees leave a company for different reasons and the organisation recruits new employees within a certain period of time to replace those who left the company (Ongori, 2007). Alao&Adeyemo (2013) describes employee attrition as the reduction or loss of employees through diverse conditions. Employee Attrition is one of the key challenges bedevilingorganizations across the globe. Employee attrition has indeed been viewed as a crucial concern for businesses due to the negative impact it has on workplace motivation and productivity as well as prolonged growth strategies. It is difficult to predict and sometimes inject observable void in an organization's skilled workforce (Alao&Adeyemo, 2013). Sexton et al (2005) reiterate that attrition's effects on service company is devastating as it distorts timely delivery of services, decreases overall firm productivity and also reduces customer loyalty whenever employee leaves abruptly. Tanmay (2018) in his work, identified some of the factors which can trigger employees to leave a company, such as; salary dissatisfaction, stagnant career growth, etc.

In the light of the above, it is crucial for companies to devise sound recruitment, acquisition and retention policy and implement effective mechanism to prevent employee attrition. Shenghuan &Tariqul (2021) opined that for an organisation to maintain higher competitive advantage over its competitor, it should ensure to reduce employee attrition.

In recent years, there has been a massive advancement in Artificial Intelligence (AI) with the development of robotics, computer vision, machine learning, deep neural networks and natural language processing. "Artificial Intelligence is the science and engineering of making intelligent machines; especially intelligent computer programs" (Mitchell, 1999). This has led to development of powerful quantitative techniques to derive insights from industry data (Zhao et al, 2018).

Employee Attrition Prediction helps Human resources (HR)Managers to predict how many employees will resign from the company during a particular period. Thus, Managers can work on deciding on the valuable employees and try to retain them. Employee Attrition Prediction and automation of the retention decision can be done by applying supervised training method and various classification algorithms (Kelleher, 2017). "Supervised machine learning is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances" (Kotsiantis, 2007).

Organizations can leverage on machine learning (ML) algorithms to get insight into their data by predicting the employee's future scope (Suvoj& Lydia, 2022).The outcome can be used by the company's HR department to cater for the existing employees which is better than choosing employees during the next recruitment process .ML is a subfield of artificial intelligence .ML describes an idea of a computer program learning and adapting to new data without the need for human interference. Supervised learning, a prominent technique in machine learning has been used extensively to analyze employee attrition problem.

Nagadevara et al (2008) described the performance evaluation of the different supervised algorithm in analyzing employee attrition. Ajit(2016) provided a compressive comparison of the predictive capabilities of seven machine learning algorithms on employee turnover. Similarly, sikaroudi et al (2015) in their work used ten different data mining algorithms to predict employee turnover including tests on different neural networks and induction techniques. Additionally, several researchers have also made progress in their quest to better understand which features (e.g. age, work experience, salary etc.) are most influential in predicting employee attrition. Despite the depth of the research outcome described above, the prediction of employee attrition that emanate from machine learning methods are seldom problem –specific and difficult to generalize (Zhao et al,2018).

In this paper, we presented an evaluation of supervised machine learning algorithms such as decision tree, randomforest classifier, logistic regression and XGBoost on an open-source (IBM HR) data set. The main focus is to predict whether the employees will leave their current companies within the specific time interval. It is formulated as a binary classification problem which classifies employees as who "will leave" (turnover) and who "will not leave" (no-movement). Succinctly, we provided comprehensive information on each supervised machine learning methods, benefits, capabilities in the context of predicting employee attrition. Importance of sampling, feature engineering and data visualization were also discussed.

Employee Attrition

Employee attrition is a costly problem faced by many employers, albeit to different degrees. Employee attrition is described as the natural process by which employee exit from the workforce - for instance, through resignation for personal reason or retirement and are not replaced on time (Budhwar&Bhatnargar, 2007). It is also known as Employee Turnover or Employee defection. Purohit (2016) described Employee turnover as the number of employees who are exiting from the company over a period of time and it is generally expressed in the percentage of the total number of employees in an organization. Furthermore, staff turnover is the rate at which employees leave and replaced by others in a Company.

Machine Learning

Machine learning is a subset of Artificial intelligence (AI) , where computers are taught to learn from data to solve problems, this is known as AI . Data is used to extract knowledge, which is then used by machine learning (ML) algorithms to create precise predictions or enhance performance (Naqa& Murphy,2015). The term "ML" also refers to predictive analysis.

Supervised Machine Learning Algorithms for Classification

K-Nearest Neighbor (kNN)

The K-Nearest Neighbor (kNN) algorithm is one of the supervised machine learning algorithms. It is utilized in pattern recognition and data mining for classification with a low error rate. The k samples that are most closely similar to a query is found by the algorithm. The category of the query point q is the same as the category of most of these examples

Support Vector Machine (SVM)

Support Vector Machine (SVM) is an effective classification algorithm. The algorithm draws a vector between the two classes on the plane at the farthest distance from both classes to separate two classes. It is widely used for classification. Figure 2.0 shows the working principle of the SVM.

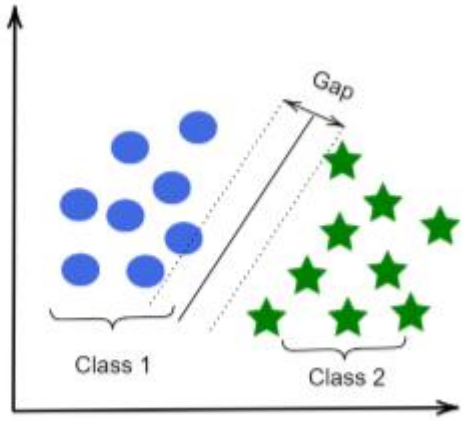


Fig 1.0

Random Forest (RF)

Random forest, a tree-based algorithm, is well known in machine learning problems. Random Forest (RF) is utilized for classification problems. The random forest that works by producing multiple decision trees generates multiple random training subsets. Then it creates a tree with random training subsets.

Logistic Regression

Logistic regression is a specific case of linear regression models. LR generates basic possibility classification formulas by using the maximum likelihood ratio when producing the equation. LR shows the statistical importance of the variables. LR is beneficial in status which the dependent variable is a dichotomy. Figure 3.0 shows the working principle of Logistic Regression.

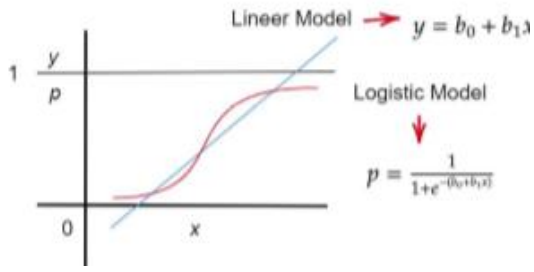


Figure 2.0

XGBoost

XGBoost is an optimized distributed gradient boosting. It is a library intended to be highly efficient, flexible, and removable. XGBoost provides a parallel tree boosting to solve many data science problems. XGBoost that adopts the

principle of gradient boosting, which is a boosted tree algorithm. XGBoost utilizes a more regularized-model to control over-fitting then other gradient boosted machines. The function contains each the structure of the tree and the leaf scores. This is formalized as:

$$f_t(x) = w_{q(x)}\, w \in R^T, q: R^d \rightarrow \{1,2,\cdots,T\}$$

1 LITERATURE REVIEW

Suvoj& Lydia (2022) carried out a comparison study on the application of supervised machine learning algorithms like Random Forest Classifier, Logistic Regression, Naive Bayes Classifier AND XGBoost to predict employee's attrition in a company. They used data set from Kaggle which consists of information on employees at a company. The model developed were evaluated on the test set using the different evaluation measures to determine the most efficient model. The result showed that XGBoost classifier has better mean accuracy 85.3%, followed by the Random Forest classifier with 83% accuracy than the other two algorithms.

Norrman (2020) studied the application of supervised ML algorithms on the prediction of employee turnover on an individual level and its impact on an organisation. The author concluded that neither of the techniques could predict employee turnover with the tested data sets.

Chourey et al (2019) in their research work carried out a survey on employee attrition prediction using Machine learning models. The authors emphasize attrition as major problem in every organisation. They made a case for algorithms for predicting employee turnover using data mining techniques

Lucas (2109) developed a framework to correctly study the problem of employee churn for a particular organisation. The developed framework entails understanding the population and many sub-groups in the data and their relationship with churn; developing a classification model for employee attrition, investigating the main reason that encourage decision to leave the company. The research work develops an extended generic method, throughautoencoders and time-series embeddings to thoroughly identify the numerous types of churning employee and categorize them in one of a three group

Jain &Nayyar (2018) applied extreme gradient boost (XGboost) ML technique to predict employee attrition from an online data set. The model designed using XGboost has more than 50% specificity rate and less than 30% error rate, also the accuracy almost reached 90%. They recommended that XGboost technique as a top priority for employee turnover prediction.

2 METHODOLOGY

This research follows experimentation research method, where data can be measured using statistical analysis. The dataset was analyzed to detect the most important features that affect the prediction and builds a predictive model according to the following procedure.

i. Gathering employees' data: IBM dataset has been used.
ii. Pre-processing the collected data: Data are prepared to be utilized by the predictive model.
iii. Analyzing the dataset: The most important features that push an employee to leave the organization are detected.
iv. Balancing the dataset: Since the dataset is not already balanced, it is necessary to be equalized.
v. Building the predictive model: The suitable con□guration for the model is selected to increase the prediction accuracy.
vi. Validating the model: K-fold validation and **70%:30%** train-test set is used for system evaluation.

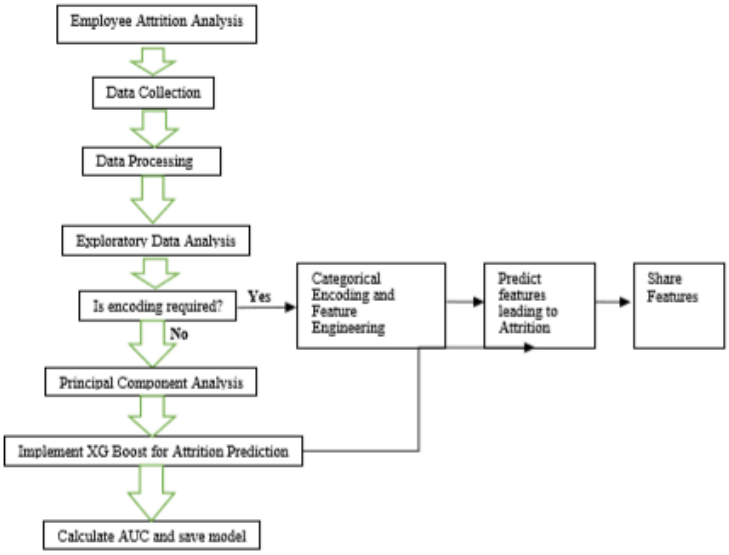


Fig. 3.0 Proposed Architecture

Data Collection and Processing

Data Collection and Dataset Description

The dataset used in this project was created by IBM Analytics. The data set consists of 1470 rows (instances) and 35 columns (features) with both categorical and numerical attributes. Each field on a record in the employee data set contains information about a specific employee. The dataset features along with their corresponding types are illustrated in Table 2. The "Attrition" feature represents the employee decision: Yes (leave the company) or No (stay at the company).

Table 3.0: IBM data features

Table 2. IBM dataset features.

Feature Name	Type	Feature Name	Type
Age	Number	MonthlyIncome	Number
BusinessTravel	Category	MonthlyRate	Number
DailyRate	Number	NumCompaniesWorked	Number
Department	Category	Over18	Category
DistanceFromHome	Number	OverTime	Category
Education	Category	PercentSalaryHike	Number
EducationField	Category	PerformanceRating	Number
EmployeeCount	Number	RelationshipSatisfaction	Category
EmployeeNumber	Number	StandardHours	Number
EnvironmentSatisfaction	Category	StockOptionLevel	Category
Gender	Category	TotalWorkingYears	Number
HourlyRate	Number	TrainingTimesLastYear	Number
JobInvolvement	Category	WorkLifeBalance	Category
JobLevel	Category	YearsAtCompa	Number
EducationField	Category	YearsInCurrentRole	Number
JobRole	Category	YearsSinceLastPromotion	Number
JobSatisfaction	Category	YearsWithCurrentManager	Number
MaritalStatus	Category	Attrition	Category

Exploratory Data Analysis

Importing data set

For this project, the data set is uploaded to Jupiter notebook in anaconda. The read_csv() function from pandas is used to read the CSV file. The command 'data.head' shows the following data set with the number of rows and columns. The head() method returns the first 5 rows of a record if a number is not specified.

```
In [22]: data.head()
```

Out[22]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...	RelationshipSatis
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	...	
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	...	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	...	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	...	
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	...	

5 rows × 35 columns

. Figure 3.1 Overview of Imported data set

Number of rows and Columns

The command data. shape gives the number of rows and column in the data set. In this case, we have 1470 rows and 35 columns

```
In [23]: data.shape

Out[23]: (1470, 35)
```

Figure 3.2Numbers of rows and columns in the data set

Dataset Balancing

The dataset used in this project is target biased. This implies that the number of employees that left the organization (attrition = "yes") is not equivalent to the number of still working employees (attrition = "no") as shown in Figure 3.3.

```
In [16]: total_rows = data.shape[0]
         no_val = data.Attrition.value_counts()[0]
         yes_val = data.Attrition.value_counts()[1]
         print('Percentage of NO Values: ' + str((no_val/total_rows) * 100))
         print('Percentage of YES Values: ' + str((yes_val/total_rows) * 100))

         plt.figure(figsize = (7, 7))
         plt.pie([yes_val, no_val], labels=['YES', 'NO'], autopct='%1.0f%%', colors = ['lightgreen','#66b3ff'])
         plt.title("Observed Class Imbalance in Attrition values", fontsize=20)
         plt.show()

         Percentage of NO Values: 83.87755102040816
         Percentage of YES Values: 16.122448979591837
```

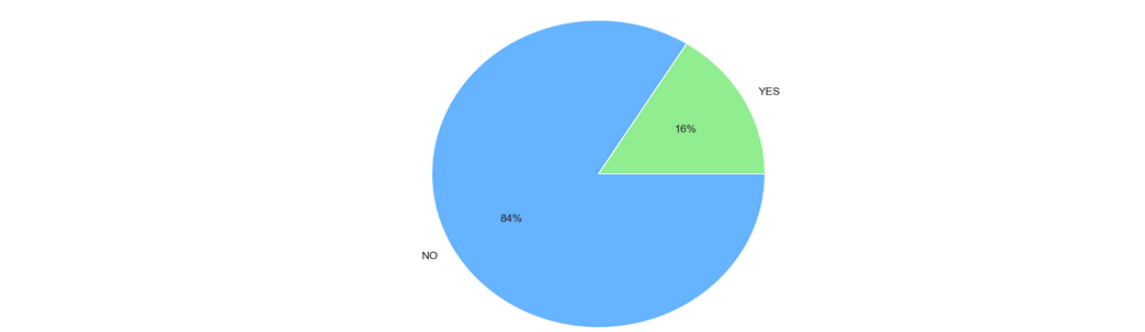


Figure 3.3 Dataset imbalance

Observation

From diagram in figure 3.3, it obvious that the data is completely imbalanced as the number of those have left is 84 while those still working is 16

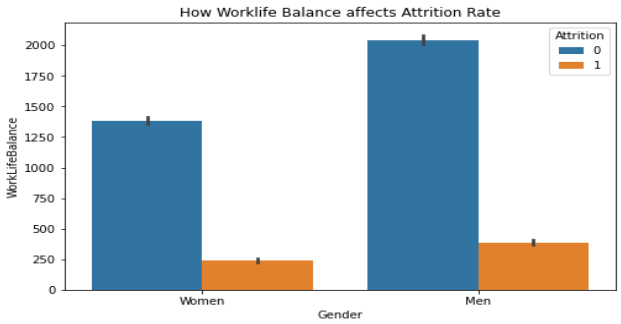


Figure 3.4

Observation

It can be deduced from the diagram that Married people have a maximum Work-Life Balance rate than others.

4. Results

Experiment 1

In this experiment, the original dataset was used, this posed a challenge due to the big difference between the number of samples of target 0 and target 1. Table 4.1 shows the performance of the different classification models. Results demonstrated that the accuracy and f1-score (0.396) of our XGBoost model outperform signi□cantly all competitor methods, mainly due to the classi□cation power of XGBoost and our pre-processing steps. The result shows that the accuracy is better in XGBoost and Logistic regression models.

Table 4.1: Classification Models Scores comparison for imbalanced data set

Out[30]:

	Name	Accuracy	Specificity	Precision	Sensitivity	ROC AUC Score	F1 Score
0	Logistic Regression	0.820862	0.991758	0.250000	0.012987	0.502373	0.024691
1	XGBoost	0.854875	0.978022	0.724138	0.272727	0.502373	0.396226
2	Decision Tree	0.777778	0.868132	0.360000	0.350649	0.609391	0.355263
3	Random Forest	0.836735	0.986264	0.666667	0.129870	0.558067	0.217391

Table 4.2: Classification Models for imbalanced data set

	Name	Mean Accuracy	Minimum Accuracy	Maximum Accuracy
0	Logistic Regression	0.664684	0.624390	0.723301
1	XGBoost	0.859072	0.843902	0.873786
2	Decision Tree	0.771594	0.741463	0.796117
3	Random Forest	0.857135	0.848780	0.868932

Experiment 2

In this experiment, the original dataset was converted into a synthetic one using SMOTE method. Table 4.3 shows the performance evaluation of results of different models under balanced data sets.Unexpectedly, results demonstrated that the accuracy and f1-score (0.507) of our XGBoost model still outperform signi□cantly all competitor methods in a balanced data set. The result shows that the accuracy is better in XGBoost and random forest models

Table 4.3: Classification Models Score comparison for balanced data set(oversampling)

Out[52]:

	Name	Accuracy	Specificity	Precision	Sensitivity	ROC AUC Score	F1 Score
0	Logistic Regression	0.684807	0.717033	0.284722	0.532468	0.624750	0.371041
1	XGBoost	0.845805	0.928571	0.573770	0.454545	0.624750	0.507246
2	Decision Tree	0.716553	0.782967	0.281818	0.402597	0.592782	0.331551
3	Random Forest	0.820862	0.942308	0.475000	0.246753	0.594530	0.324786

Table 4.4: Classification Models Scores comparison for balanced data set

Out[67]:

	Name	Mean Accuracy	Minimum Accuracy	Maximum Accuracy
0	Logistic Regression	0.664684	0.624390	0.723301
1	XGBoost	0.859072	0.843902	0.873786
2	Decision Tree	0.771594	0.741463	0.796117
3	Random Forest	0.857135	0.848780	0.868932

In [ ]:

4. Discussion

XGBoost classification algorithm has shown superiority over other machine learning algorithms in the prediction problem. In spite of the imbalanced nature of the data set, the model showed high prediction accuracy over other algorithm evaluated. The fit was not only achieved by the XGBoost algorithm, but also due to proper adoption of pre-processing and robust features selection technique. In experiment 2, synthetic balanced data set was used and the result compared to the results in the imbalanced classification. It was observed that most of the algorithms performed excellently in imbalanced data set than balanced data set.

5. Conclusion

In this research work, we have evaluated the performance of different algorithm in predicting employee attrition in an IBM data set. The outcome can assist the human resources unit as regard important information about a potential decision of an employee to leave the company. Similarly, early detection of attrition may help the employee to rectify himself and put more efforts towards his/her work responsibilities. Our method predicts the most important features that contribute to attrition. Similarly, the correlations among different features were also presented. From our finding, we discovered that monthlyincome, overtimehour, joblevel are the most significant feature that influence the

employee decision. The data set offered by IBM proves to be challenging due to its imbalance nature. This results to creating a synthetic version of the dataset to develop a classifier that can aid real life prediction. Through the experiment conducted, the efficiency of our algorithm was measured in terms of accuracy, recall, f1 score and precision. The accuracy for Logistic regression, XGBoost, decision tree and Radom Forest for imbalanced data sets are 82.01%,85.4%,77.7% and 83.6% respectively. Similarly, the accuracy for logistic regression, XGBoost, decision tree and Random Forest for synthetic balanced dataset are 68.48%, 84.58%, 71.66% and 82.08%respectively. The algorithm that produced the best result for the dataset was XGBoost Algorithm.

Thus, it can be inferred that XGBoost is a powerful algorithm that provides the best accuracy in our use case. Employees are the backbone of the organization and it is very crucial for any company to successfully retain its employee. From our study, we came to the conclusion that various factors like monthlyincome, overtimehour, joblevel are crucial in deciding the attrition rate for a company. In order to visualize the performance of this binary classifier which is built using different classification algorithms, we use the AUC-ROC metric. It summarizes the trade-off between the true positive and false positive sets. Since we have already balanced the target class using SMOTE, we can predict the probabilities using ROC (Receiver Operating Characteristic curve).

6. Future Work

In future research work, it possible to improve the analysis by considering the working condition (hazard, harm), racial profiling, religion etc. that are positively related to employee attrition intention. Similarly, we shall attempt to deploy the python file used for machine learning models on Heroku cloud infrastructure. This project can be extended in future by applying other classification algorithms and compare the result with deep learning techniques.

References

Ajit, Pankaj(2016). "Prediction of employee turnover in organizations using machine learning algorithms." algorithms 4.5 : C5.

Alao, D., Adeyemo, A.B. (2013).Analyzing employee attrition using decision tree algorithms. Comput. Inf. Syst. Dev. Inform. Allied Res. J. 4

Budhwar P S, Bhatnagar J.(2007). Talent management strategy of employee engagement in Indian ITES employees: key to retention[J]. Employee relations.

Chourey A, Phulre S, and Mishra S (2019). A survey paper on employee attrition prediction using machine learning techniques."

IBM HR Analytics Employee. Available online: https://www.kaggle.com/zainsr4/ibm-hr-analytics

Jain R and Nayyar A (2018). Predicting employee attrition using xgboost machine learning approach," in 2018 international conference on system modeling& advancement in research trends (smart). IEEE, 2018, pp. 113–120.

Karina (2020).What to Do When You Feel Underappreciated at Work - Worksion,Nov. 2020, section: Career Growth. [Online]. Available: https://worksion.com/underappreciated/

Lucas Santos Fischer (2019). A predictive and prescriptive framework for employee churn. Dissertation Submitted in partial fulfillment of the requirements for the degree of Master of Science in Analysis and Engineering of Big Data, Faculty of Sciences and Technology, NOVA University of Lisbon.

Mitchell TM (1999). Machine Learning and Data Minning. Communications of ACM .42: 31-36

Naqa El and Murphy M.J (2015). What is machine learning?" in machine learning in radiation oncology. Springer, pp. 3–11.

Nagadevara, V., Srinivasan, V., Valk, R(2008). Establishing a link between employee turnover and withdrawal behaviours: application of data mining techniques. Res. Pract. Hum. Resour.Manag. 16, 81–97

Norrman F (2020), "Predicting employee attrition with machine learning on an individual level, and the effects it could have on an organization,"

Purohit, M. (2016). A study on employee turnover in IT sector with special emphasis on Wipro and Infosys. IOSR Journal of Business and Management, 18(4), 2319–7668. doi: 10.9790/ 487X-1804014751

Ongori, H. (2007). A review of the literature on employee turnover, African Journal of Business Management, 2(1): 049-054, Department of Management, University of Botswana, Botswana, available at: http://www.academicjournals.org/ajbm, accessed on August 8 , 2022

Sexton, R.S., McMurtrey, S., Michalopoulos, J.O., Smith, A.M(2005).: Employee turnover: a neural network solution. Comput. Oper. Res. 32, 2635–2651

Sikaroudi, E., Mohammad, A., Ghousi, R., Sikaroudi, (2015)A.: A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). J. Ind. Syst. Eng. 8, 106 121

Suvoj Reddy Kovvuri& Lydia Sri Dommeti,(2022) .Employee Turnover Prediction A Comparative Study of Supervised Machine Learning Models. B.sc thesis is submitted to the Faculty of Engineering at Blekinge Institute of Technology in partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Science

Shenghuan Yang, Md Tariqul Islam ,(2021) IBM Employee Attrition Analysis,arXiv:2012.01286v6

Zhao, Y., Hryniewicki, M.K., Cheng, F., Fu, B., Zhu, X. (2019). Employee Turnover Prediction with Machine Learning: A Reliable Approach. In: Arai, K., Kapoor, S., Bhatia, R. (eds) Intelligent Systems and Applications. IntelliSys 2018. Advances in Intelligent Systems and Computing, vol 869. Springer, Cham. https://doi.org/10.1007/978-3-030-01057-7_56